OMNI-CUSTOMIZER: END-TO-END MULTIMODAL CUSTOMIZATION FOR JOINT AUDIO-VIDEO GENERATION

Anonymous authors
Paper under double-blind review

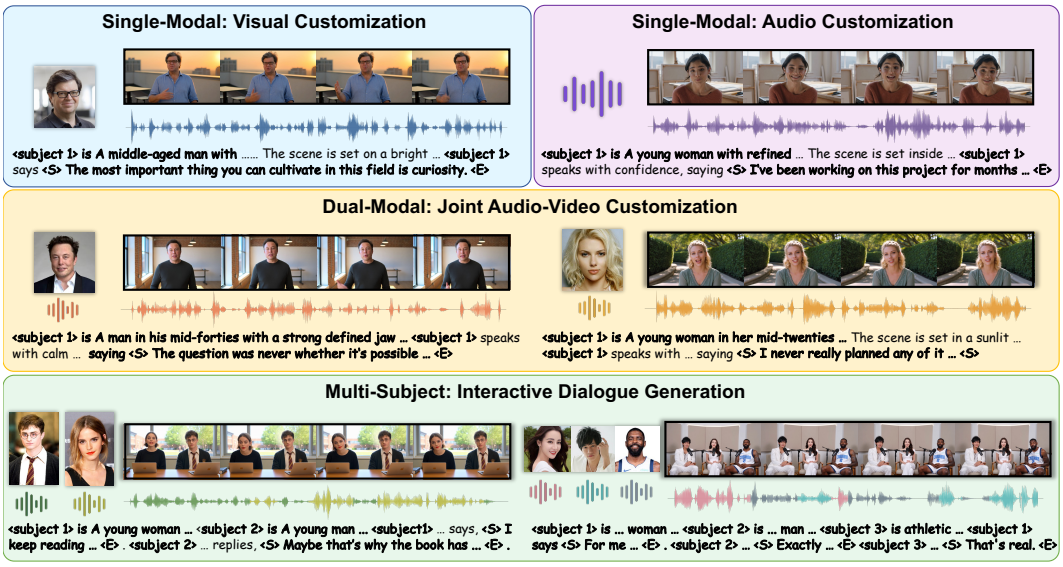


Figure 1: Omni-Customizer achieves high-quality joint audio-video customization conditioned on ***1)*** reference images, ***2)*** reference audio, or ***3)*** both. Furthermore, it demonstrates robust multimodal binding capabilities in ***4)*** highly realistic, multi-subject conversational scenarios.

ABSTRACT

Despite recent advances in joint audio-video generation, precise appearance-and-voice customization remains challenging. We present Omni-Customizer, an end-to-end framework for human-centric joint audio-video customization. To establish cross-modal identity associations, we introduce Omni-Context Fusion (OCF) to integrate text embeddings with visual and audio references, together with Semantic-Anchored Multimodal RoPE (SA-MRoPE) to anchor paired references to their corresponding subjects. To reduce face duplication and timbre leakage under incomplete reference coverage, we further propose Subject-Aware Reference Routing (SARR), which controls where and how broadly each reference can directly influence generated content. We adopt a three-stage progressive training curriculum to achieve stable optimization and robust customization. We also develop a human-centric data curation pipeline to construct high-quality multimodal training data. Experiments on our proposed OC-Bench demonstrate strong video and audio quality, robust appearance-and-voice preservation, and faithful identity binding in complex multi-subject and incomplete-reference settings.

1 INTRODUCTION

Following the open-source release of foundational models such as Ovi (Low et al., 2025) and LTX-2 (HaCohen et al., 2026), joint audio-video generation (JAVG) has attracted growing attention. Re-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

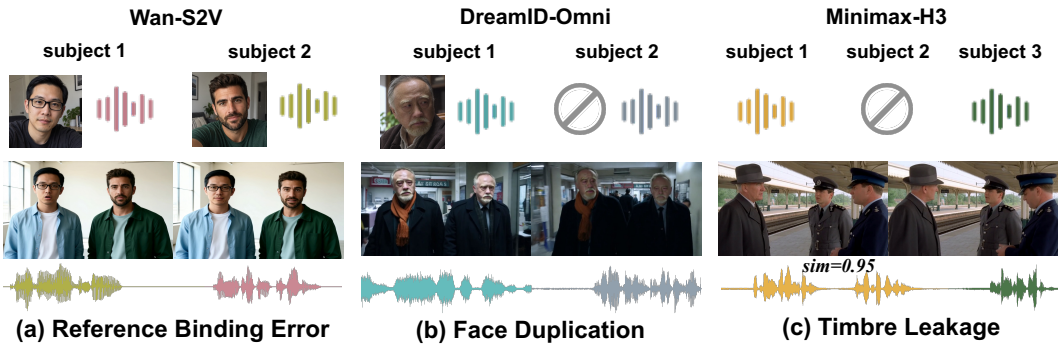


Figure 2: **Representative limitations of existing audio-video customization methods.** Existing approaches may suffer from (a) incorrect cross-modal identity binding and identity leakage under incomplete reference coverage, including (b) face duplication and (c) timbre leakage.

cent proprietary models such as Seedance 2.0 (Seedance et al., 2026) further demonstrate the broader application potential of JAVG, such as cinematic content creation, digital human generation, and interactive media. Despite these advances, precise and robust multimodal customization for JAVG remains underexplored in open-source research, particularly in realistic and complex generation settings that require coherent control over multiple subjects and heterogeneous cross-modal references.

Existing customization efforts can be broadly organized into two paradigms. ***1) Decoupled customization pipelines.*** Visual personalization (Liu et al., 2025b; Li et al., 2025b; Jiang et al., 2025) and audio-driven generation (Chen et al., 2025b; Gao et al., 2025) provide complementary capabilities that can be composed for personalized audio-video generation. However, appearance and voice conditions are introduced through separate components or stages, without an explicit mechanism that jointly binds a reference image and a reference audio to the same identity. ***2) Joint customization.*** Unified generators associate image and audio references through structured captions and positional indexing (Guo et al., 2026), in-context adaptation (Dahan et al., 2026), shared identity embeddings (Chen et al., 2026a), or temporal reference masks (Liu et al., 2026). Yet incomplete reference coverage remains challenging when some characters have no references or only one modality of reference is available, as the provided identity conditions may be over-applied to unmatched subjects, creating a risk of face duplication or timbre leakage.

To overcome these limitations, we propose ***Omni-Customizer***, an end-to-end framework for human-centric joint audio-video customization. To establish precise cross-modal identity associations, we first introduce **Omni-Context Fusion (OCF)**, which extends text conditioning with visual and audio reference information, together with **Semantic-Anchored Multimodal RoPE (SA-MRoPE)**, which anchors the references to their corresponding subjects while preserving their respective modality-specific positional structures. To further prevent the available identity references from being over-applied to other characters under incomplete reference coverage, we propose **Subject-Aware Reference Routing (SARR)**, which selectively controls where and how broadly each reference can directly influence the generated tokens. Finally, to improve training stability and robustness under complex multi-subject and incomplete-reference settings, we adopt a **progressive training strategy** that advances from single-subject customization to multi-subject generation and subsequently introduces reference dropout and consistent subject reindexing. Together, these designs enable robust appearance-and-voice customization in complex multi-subject scenarios, including those with incomplete reference coverage, as shown in Fig. 1.

Extensive experiments on our newly proposed Omni-Customizer Benchmark (**OC-Bench**) demonstrate the effectiveness of our framework. Omni-Customizer achieves strong video and audio quality and accurate audio-video synchronization. Furthermore, it achieves robust appearance-and-voice customization with reliable cross-modal identity binding even in complex multi-subject scenarios, including those with incomplete reference coverage, thereby validating the effectiveness of our architectural designs and training strategy. In summary, our main contributions are as follows:

1) We propose Omni-Customizer, an end-to-end framework tailored for human-centric joint audio-video customization in complex multi-subject scenarios with heterogeneous visual and audio references, including those with incomplete reference coverage.

2) We introduce Omni-Context Fusion (OCF) together with Semantic-Anchored Multimodal RoPE (SA-MRoPE) to establish precise cross-modal associations between paired visual and audio identity references while preserving their modality-specific positional structures.

3) We propose Subject-Aware Reference Routing (SARR), which explicitly controls where and how broadly each identity reference can influence the generated content, reducing face duplication and timbre leakage under incomplete reference coverage.

4) We develop a progressive training strategy that gradually extends customization from single-subject to multi-subject settings and further incorporates reference dropout and consistent subject reindexing to improve training stability and robustness.

5) We develop a human-centric multimodal data curation pipeline and construct OC-Bench, covering both standard joint-customization settings and challenging partial-modality and partial-subject reference conditions for comprehensive evaluation.

2 Multimodal Data Curation

Data Sources and Filtering. We curate a multi-subject audio–video customization dataset from OpenHumanVid (Li et al., 2025a), OpenS2V-5M (Yuan et al., 2025), TalkVid (Chen et al., 2026b), and our in-house video collection. We use the quality scores available in the source metadata to retain higher-quality videos. We further retain clips with a resolution of at least 480p, a frame rate of at least 24 FPS, and a duration of 3–10 seconds.

Reference Image Construction. We sample frames at 1 FPS from the full unfiltered video collection and perform face detection using InsightFace (Guo et al., 2018). Detected faces are clustered across clips according to ArcFace (Deng et al., 2019) similarity. Within each identity cluster, we first filter low-quality images using DSL-FIQA (Chen et al., 2024), followed by a further quality review with Qwen3.5-397B-A17B (Qwen Team, 2026). To construct cross-pair references that preserve identity while minimizing shared scene and contextual information with the target clip, we favor candidate reference images with high ArcFace similarity but relatively low full-frame similarity measured by Qwen3-VL-Embedding-8B (Li et al., 2026b). *Each selected reference image is resized while preserving its aspect ratio so that its long side matches that of the target video, with equal white padding added to both sides along the short dimension.*

Transcription and Multimodal Annotation. We obtain timestamped transcripts using Qwen3-ASR and Qwen3-ForcedAligner (Shi et al., 2026b). Gemini 2.5 Pro (Comanici et al., 2025) receives each audio–video clip, its reference images, and the timestamped transcript to produce a structured caption. The caption links subjects to their reference images and spoken utterances while describing visual actions and acoustic context, with the detailed structure explained in Sec. 3.1.

Reference Audio Construction. Using the structured captions and aligned transcripts, we select suitable speech segments and their transcriptions for each subject as voice-cloning conditions. We generate multiple reference audios using several voice-cloning models (Xin et al., 2026; Lian et al., 2026; Du et al., 2025). For each subject, we select the generated reference audio with the highest WavLM-Large (Chen et al., 2022) similarity to the corresponding source speech and discard samples that fall below the acceptance threshold. *All selected reference audios are cropped to 5 seconds.*

Overall, this pipeline yields a 1M-scale human-centric audio–video customization dataset. Further implementation details of the curation pipeline and additional statistics of the curated dataset are provided in the supplementary material.

3 Method

3.1 Formulation of Joint Audio-Video Customization

Reference-Conditioned Dual-Stream Generation. Omni-Customizer is built upon a dual-stream Diffusion Transformer (DiT) initialized from the pre-trained Ovi (Low et al., 2025) backbone. For N subjects, the inputs comprise reference images $\mathcal{I} = \{I_k\}_{k=1}^{N}$, reference audios $\mathcal{A} = \{A_k\}_{k=1}^{N}$, and a text prompt P. A missing image reference is replaced with a pure white image, and a missing audio reference with a silent audio clip. The corresponding VAEs encode each subject's references as c_v^k and c_a^k, while the text encoder maps P to c_{txt}. At denoising timestep t, the noisy target video

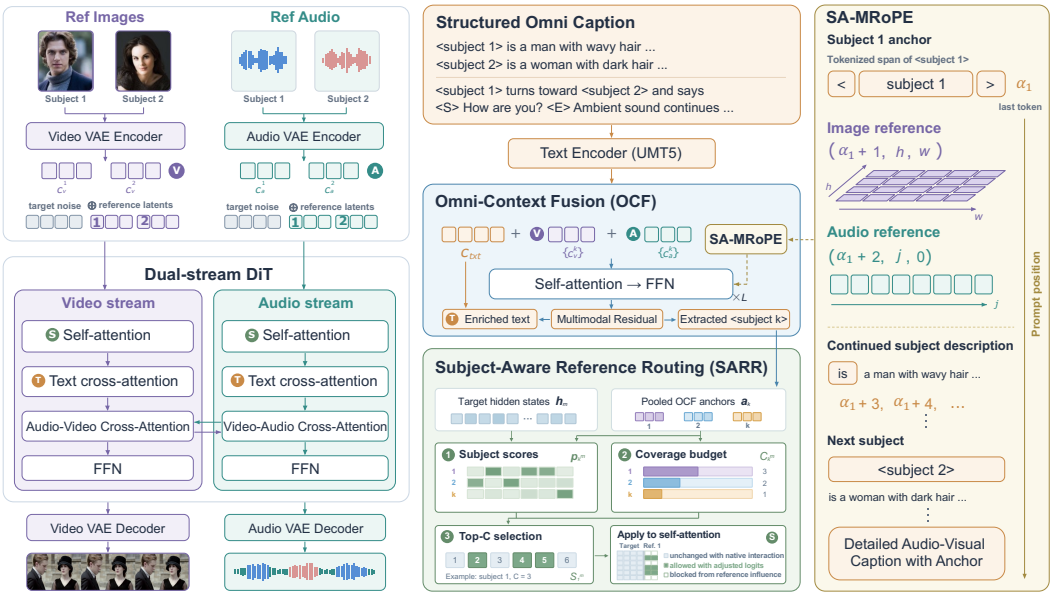


Figure 3: Framework of Omni-Customizer: The text prompt, reference images, and audios are integrated by OCF, with SA-MRoPE providing subject-level semantic anchoring. SARR further controls where and how broadly each reference directly influences target tokens. The resulting conditions guide the dual-stream backbone for robust multi-subject appearance-and-voice customization.

and audio latents, denoted by $z_{v,t}$ and $z_{a,t}$, are jointly processed as

$$(\hat{y}_v, \hat{y}_a) = \mathcal{F}_\theta\Big([z_{v,t} \oplus c_v^1 \oplus \cdots \oplus c_v^N], [z_{a,t} \oplus c_a^1 \oplus \cdots \oplus c_a^N], t, c_{\text{txt}}\Big) \tag{1}$$

where $\oplus$ denotes sequence concatenation, and $\hat{y}_v$ and $\hat{y}_a$ are the video and audio predictions. Let T_v and T_a denote the temporal lengths of the target latents, and let R_v and R_a denote those of visual and audio reference latents. Using zero-based indexing, the temporal RoPE indices are $T_v + (k-1)R_v, \ldots, T_v + kR_v - 1$ for c_v^k and $T_a + (k-1)R_a, \ldots, T_a + kR_a - 1$ for c_a^k. On the remaining dimensions, each reference uses the same RoPE indices as the corresponding target latents.

Structured Omni-Caption. The structured prompt consists of subject-specific multimodal descriptions followed by a detailed audiovisual description, where the defined anchor token is consistently used to refer to each subject. Formally, the prompt P is structured as:

$$\underbrace{\texttt{<subject 1> is } D_{v,1}\texttt{, with } D_{a,1}\texttt{.}}_{P_{sub,1}\text{ Subject 1 Description}} \ldots \underbrace{\texttt{<subject N> is } D_{v,N}\texttt{, with } D_{a,N}\texttt{.}}_{P_{sub,N}\text{ Subject N Description}}$$
$$\underbrace{D_{\text{env}}\ D_{\text{act}}(\texttt{<subject } i\texttt{> acts...}) \ldots \texttt{<subject } k\texttt{> says <S> } T_{k,j}\texttt{ <E>.}\ D_{\text{sound}}}_{P_{\text{detail}}\text{ Detailed Audiovisual Description}} \tag{2}$$

Here, $D_{v,i}$ and $D_{a,i}$ denote the visual and acoustic descriptions of subject i, respectively. The subsequent P_{detail} follows a natural detailed audiovisual description, typically beginning with the scene and environment D_{env} and then describing subject actions and interactions D_{act}, speech, and background or other non-speech sounds D_{sound} in temporal order.

3.2 Omni-Context Fusion and Semantic Anchoring

Relying solely on textual prompts to associate multimodal identity conditions $\{c_v^k, c_a^k\}$ with the correct spatiotemporal regions can be unreliable in multi-subject scenes. In the backbone, visual and audio references are introduced through separate modality streams without explicit subject-level cross-modal binding cues. To achieve more precise cross-modal identity binding, we therefore allow text, visual and audio references to interact directly in a dedicated condition module.

Omni-Context Fusion (OCF). We introduce OCF to extend text-only conditioning with multimodal encoding, incorporating visual and audio identity references into the prompt embeddings. Specifically, we concatenate the text embeddings c_{txt} with all visual and audio reference tokens to form $S = [c_{\text{txt}} \oplus c_v^1 \oplus \cdots \oplus c_v^N \oplus c_a^1 \oplus \cdots \oplus c_a^N]$. The sequence is processed by L Transformer blocks to jointly encode text and reference information. After the final block, we extract the states at the text positions and inject their difference from the input text states as a residual update to c_{txt} (He et al., 2016). The output projections of the attention and feed-forward branches are zero-initialized, so OCF is an identity mapping at initialization. Only this text residual is returned to the denoising backbone; the visual and audio reference outputs of OCF are not written into the backbone residual stream. The video and audio streams query the enriched text embeddings through their respective text cross-attention layers, accessing a shared representation of appearance and voice during denoising. Together with the subject anchoring provided by SA-MRoPE (described next), this shared conditioning provides explicit cues for associating each subject's visual and vocal identities.

Semantic Anchored Multimodal RoPE (SA-MRoPE). While OCF aggregates multimodal inputs into a unified sequence, their standard sequence positions do not explicitly associate each reference with its corresponding subject description. Moreover, text tokens follow a one-dimensional sequence, whereas visual and audio references possess intrinsic spatial and temporal structures, respectively. We therefore introduce SA-MRoPE to anchor each reference to its corresponding subject anchor token in the prompt while preserving the intrinsic positional structure of each modality. Let α_k denote the temporal index of the last token of <subject k> within $P_{sub,k}$, which may be tokenized into multiple text tokens. For a visual reference token $c_{v,h,w}^k$ at spatial location (h, w) and an audio reference token $c_{a,j}^k$ at temporal index j, we assign

$$Pos(c_{v,h,w}^k) = (\alpha_k + 1, h, w), \qquad Pos(c_{a,j}^k) = (\alpha_k + 2, j, 0). \tag{3}$$

The subsequent text tokens resume from temporal position $\alpha_k + 3$. In this way, SA-MRoPE facilitates subject-level cross-modal binding through aligned positional cues while preserving the intrinsic spatial and temporal structures of the visual and audio references.

3.3 Subject-Aware Reference Routing

Under incomplete reference coverage, OCF and SA-MRoPE associate the available visual and audio references with their subjects, but do not explicitly restrict their influence on subjects lacking the corresponding references. We therefore introduce Subject-Aware Reference Routing (SARR) to control the scope of reference conditioning and reduce face duplication and timbre leakage.

Subject-Anchored Scoring. We average the OCF-enhanced embeddings corresponding to <subject k> in $P_{sub,k}$ and project them with W_{anc} into the DiT hidden space to obtain the subject representation $\mathbf{a}_k \in \mathbb{R}^d$, where d is the DiT hidden dimension. Let $\mathbf{h}_m \in \mathbb{R}^{N_m \times d}$ denote the target hidden states entering the current DiT block, where N_m is the number of target tokens in stream $m \in \{v, a\}$. SARR predicts the subject relevance over all target positions as $\mathbf{p}_k^m = \text{sigmoid}(\mathbf{h}_m \mathbf{a}_k / \sqrt{d})$, where $\mathbf{p}_k^m \in [0, 1]^{N_m}$ gives the relevance of each target token to subject k. We additionally supervise the routing scores with a binary cross-entropy objective, using the annotated facial or speech regions of subject k as positives and those of other subjects as negatives.

Reference Coverage. To constrain the spatial or temporal extent of each reference, we assign it a subject-conditioned token budget that limits how many target tokens it can directly influence (Zhou et al., 2022). Let N_m denote the number of target tokens in stream m. Let $r_k^m \in \{0, 1\}$ indicate whether subject k has an actual reference in stream m. We predict the budget as $C_k^m = r_k^m \lceil N_m \, \text{sigmoid}(W_c^m \mathbf{a}_k) \rceil$, where $W_c^m \in \mathbb{R}^{1 \times d}$ is a learnable projection, rather than inferring availability from a white-image or silent-audio placeholder. The predicted coverage fraction is additionally supervised with an ℓ_1 objective against the annotated fraction of facial or speech tokens belonging to subject k when $r_k^m = 1$. We select the top-C_k^m target tokens with the highest routing scores $\mathbf{p}_k^m$ as the accessible region S_k^m for that reference.

Coverage-Constrained Reference Attention. Given the accessible region S_k^m, SARR modifies only the self-attention logits from target tokens to the reference tokens of subject k. For a target

query u and a reference key $j \in c_m^k$, the original attention logit $\ell_{u,j}^m$ is modified as

$$\tilde{\ell}_{u,j}^m = \begin{cases} \ell_{u,j}^m + \log p_{u,k}^m, & r_k^m = 1 \text{ and } u \in S_k^m, \\ -\infty, & \text{otherwise.} \end{cases} \tag{4}$$

Thus, each reference can be directly accessed by at most C_k^m target tokens in each self-attention layer, while the native self-attention among target tokens remains unchanged. We leave text and audio-video cross-attention unchanged to retain OCF-enhanced conditioning and the backbone's native cross-modal interactions. With OCF already establishing the cross-modal correspondence between the paired visual and audio references, SARR independently routes their influence in the two generation streams. Further implementation details are provided in Appendix C.

3.4 Progressive Training Strategy

To promote training stability and robust customization, we adopt a three-stage curriculum optimized with the flow matching objective (Lipman et al., 2022; Liu et al., 2022).

Stage 1: Single-Subject Adaptation. We first train the model on single-subject samples with paired image and audio references. During this stage, the model learns basic appearance and voice customization in a simpler setting, facilitating stable initial adaptation.

Stage 2: Multi-Subject Customization. We then train on multi-subject samples with paired image and audio references for every subject. This stage helps the model generalize its single-subject customization capabilities to more complex scenes involving multiple subjects.

Stage 3: Reference Dropout and Subject Reindexing. We continue multi-subject training with reference dropout and subject reindexing. For each subject, we independently drop its image and audio references with probability p, such that both modalities are simultaneously dropped with probability p^2. Dropped references are marked as unavailable and replaced with the predefined placeholders. We further randomly reindex and shuffle the subjects to reduce potential subject-ordering biases. For example, the captioner may preferentially assign `<subject 1>` to a larger or more salient subject (Abbasi et al., 2025), whose reference is also assigned the RoPE indices closest to the target latents. Benefiting from the structured organization of our annotations and the shared subject anchors, we simply shuffle the order of $P_{sub,k}$, consistently remap all subject indices throughout the prompt, and reorder the corresponding reference inputs accordingly.

4 Experiments

4.1 Experimental Details

Implementation Details. We initialize Omni-Customizer from the pre-trained Ovi backbone (Low et al., 2025) and jointly train the DiT backbone and newly introduced modules for one epoch, corresponding to approximately 16K optimization steps with a batch size of 64. We train on 3–10-second clips with resolutions ranging from 480p to 720p, using 24 FPS video and audio resampled to 16 kHz.The three stages comprise 6K, 5K, and 5K steps, respectively, with the reference-dropout probability set to $p = 0.2$ in Stage 3. The hidden dimension is $d = 3072$ for both the video and audio streams, and OCF contains $L = 8$ Transformer blocks. In addition to the standard flow-matching objective, SARR is supervised by additional objectives described in Sec. 3.3. We optimize the model using AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and a weight decay of 0.01. For the newly introduced modules, we use a cosine learning-rate schedule that decays from 1×10^{-4} to 1×10^{-5}. More implementation and inference details are provided in Appendix C.

OC-Bench To evaluate multimodal customization, we construct the Omni-Customizer **Bench**mark (**OC-Bench**) with 200 test cases in three categories. ***1) Full-Reference Customization.*** We adopt the 100-case R2AV split of IDBench-Omni (Guo et al., 2026), comprising 60 single-subject and 40 multi-subject cases with paired image and audio references for each target subject. ***2) Partial-Modality Customization.*** We construct 50 multi-subject cases in which at least one subject lacks an image or audio reference. ***3) Partial-Subject Customization.*** We construct 50 multi-character cases in which paired image and audio references are provided for only a subset of characters, while the remaining characters are specified solely through the prompt. All newly constructed benchmark



Figure 4: Qualitative comparison with state-of-the-art baselines of four different paradigms.

cases are strictly disjoint from the training set, follow the structured prompt format described in Sec. 3.1, and the remaining input conditions are manually verified for quality and correctness.

Evaluation Metrics. We evaluate the generated outputs using the following metrics. ***1) Identity Preservation.*** We measure Face Similarity and temporal Face Consistency using ArcFace (Deng et al., 2019), and Timbre Similarity (T-Sim) using WavLM (Chen et al., 2022). ***2) Audiovisual Consistency.*** We evaluate lip synchronization using Sync-C and Sync-D (Chung & Zisserman, 2016), and audiovisual semantic consistency using ImageBind (IB-Score) (Girdhar et al., 2023). ***3) Video Quality.*** We report aesthetic quality using Aesthetic-v2.5 (Schuhmann et al., 2022), imaging quality using MUSIQ (Ke et al., 2021), and temporal flickering following VBench (Huang et al., 2024). ***4) Audio Quality.*** We report Production Quality (PQ) from Audiobox Aesthetics (Vyas et al., 2023) and Word Error Rate (WER) using Whisper-large-v3 (Radford et al., 2023). ***5) Identity Binding Reliability.*** We evaluate identity binding from two complementary aspects. *Binding Accuracy (BA)* measures the fraction of eligible utterances whose voice and visible speaking face match the intended subject. We locate utterances using Qwen3-ASR and forced alignment (Shi et al., 2026b), detect active speakers with TalkNet (Tao et al., 2021), and verify face and voice identities using ArcFace and WavLM. *Identity Leakage Rate (ILR)* measures how often a character without a reference in the evaluated modality matches another character's reference identity, reported separately as Visual-ILR (V-ILR) and Audio-ILR (A-ILR). We compute V-ILR and A-ILR by matching generated face tracks and speech segments against the available reference identities using ArcFace and WavLM, respectively. More implementation details are provided in the supplementary material.

4.2 Comparisons and Analysis

To comprehensively evaluate Omni-Customizer, we compare it against leading state-of-the-art models on OC-Bench across four distinct paradigms: ***1)*** Video Customization, including Phantom (Liu et al., 2025b) and VACE (Jiang et al., 2025). We exclusively evaluate visual customization and identity preservation as these models lack native audio-generation capabilities. ***2)*** Audio-Driven Video Customization, including Humo (Chen et al., 2025b), HunyuanCustom (Hu et al., 2025b), Wan2.2-S2V (Gao et al., 2025) and SkyReel-A2 (Fei et al., 2025). We evaluate video quality and AV-sync but omit audio metrics, as the driving audio is a fixed input condition rather than a generative output. ***3)*** Qwen-Image + JAVG Models. We generate the first frame using Qwen-Image (Wu et al., 2025), and then baselines (Ovi (Low et al., 2025), LTX2.3 (HaCohen et al., 2026), Universe (Wang et al., 2025), and MOVA (OpenMOSS Team et al., 2026)) generate the video in an I2V manner. ***4)*** Joint Audio-Video Customization. Evaluates end-to-end unified multimodal customization, in-

Table 1: Quantitative comparison with state-of-the-art methods on OC-Bench. **Bold** and underline represent the best and second-best results, respectively.

Method	Identity Preservation		AV-Sync		Video Quality			Audio Quality			Binding Reliability	
	Face-Sim↑/Cons↑	T-Sim↑	Sync-C↑/D↓	IB-S↑	AQ↑	IQ↑	TF↑	WER↓	PQ↑	IB-A↑	BA↑	V/A-ILR↓
Phantom	0.658 / 0.883	-	- / -	-	0.323	0.432	0.854	-	-	-	-	0.382 / -
VACE	0.675 / 0.896	-	- / -	-	0.346	0.535	0.863	-	-	-	-	0.368 / -
Humo	0.709 / 0.942	-	3.425 / 10.22	0.125	0.523	0.614	0.888	-	-	-	0.54	0.335 / -
HunyuanCustom	0.733 / 0.955	-	3.755 / 9.839	0.182	0.575	0.653	0.909	-	-	-	0.58	0.306 / -
Wan2.2-S2V	0.775 / 0.964	-	5.868 / 8.518	0.123	0.519	0.643	0.955	-	-	-	0.63	0.283 / -
SkyReel-A2	0.762 / 0.959	-	4.221 / 9.120	0.185	0.553	0.639	0.942	-	-	-	0.61	0.292 / -
Universe-1	0.643 / 0.913	-	5.015 / 9.418	0.076	0.413	0.575	0.843	0.432	3.42	0.073	-	- / -
Ovi	0.693 / 0.935	-	5.424 / 8.939	0.085	0.436	0.593	0.865	0.343	3.65	0.085	-	- / -
MOVA	0.696 / 0.937	-	5.428 / 8.935	0.086	0.439	0.595	0.867	0.339	3.66	0.087	-	- / -
LTX2.3	0.743 / 0.953	-	6.031 / 8.210	0.093	0.485	**0.673**	0.879	0.225	3.93	0.099	-	- / -
MiniMax-H3	**0.821 / 0.982**	0.485	**6.251 / 7.795**	0.191	0.587	0.651	0.962	**0.141**	4.25	**0.131**	0.76	0.281 / 0.339
DreamID-Omni	0.790 / 0.968	0.472	6.085 / 8.020	0.189	0.585	0.649	0.946	0.285	4.13	0.113	0.75	0.285 / 0.342
Omni-Customizer	0.813 / 0.977	**0.515**	6.238 / 7.818	**0.195**	**0.593**	0.655	**0.969**	0.155	**4.33**	0.125	**0.83**	**0.265 / 0.327**

cluding MiniMax-H3 (MiniMax, 2026) and DreamID-Omni (Guo et al., 2026). All methods receive the same structured prompts, with DreamID-Omni using only a format-adapted version to match its required prompt structure without changing the underlying semantics.

Quantitative Analysis. As shown in Tab. 1, Omni-Customizer achieves strong overall performance across generation quality, identity preservation, and binding reliability. While video-only methods and cascaded pipelines (e.g., LTX2.3) maintain competitive video quality, they do not natively support joint appearance-and-voice customization. Our model achieves the highest T-Sim and BA together with the lowest face and voice identity leakage, demonstrating robust appearance–voice preservation and more reliable identity assignment. Meanwhile, it remains competitive in visual fidelity, audio-video synchronization, and speech quality, showing a favorable balance between generation quality and identity control in complex multi-subject and incomplete-reference settings.

Qualitative Analysis. As illustrated in Fig. 4, we compare Omni-Customizer with state-of-the-art baselines. Phantom exhibits facial rigidity in two-subject scenarios. LTX2.3 suffers from gradual identity drift in subsequent frames. HuMo struggles with identity preservation in dual-person customization, showing mediocre consistency. DreamID-Omni performs suboptimally in both visual and acoustic modalities, resulting in noticeable identity entanglement and drift. In contrast, Omni-Customizer achieves high-fidelity customization across both visual and acoustic modalities. Our model maintains robust identity binding and stable multi-subject consistency even in complex scenes, ensuring precise lip-sync without identity confusion.

Computational efficiency analysis. Under the two-subject 480p setting with 121 frames, our newly introduced modules incur only 3.57% additional FLOPs (6.39 TFLOPs), 3.37% additional latency (16.83 ms), and 7.91% additional peak memory (1.88 GiB) on a single NVIDIA H200 GPU relative to one Ovi denoising step.

4.3 Ablation Study.

Tab. 2 evaluates the contribution of each proposed component through both progressive and full-context ablations. In the first block, we progressively introduce OCF, SA-MRoPE, SARR scoring, and the coverage constraint without progressive training, while the second block removes SA-MRoPE, SARR, or coverage from the full training configuration to further isolate their effects. Adding OCF improves both identity preservation and BA, while SA-MRoPE further strengthens subject-level association, increasing BA from 0.63 to 0.68 and reducing V/A-ILR from 0.584/0.482 to 0.512/0.415. SARR then provides more selective reference use: under the full training setting, introducing subject scoring increases BA from 0.76 to 0.81 and reduces V/A-ILR from 0.458/0.425 to 0.324/0.354. The coverage constraint further improves BA to 0.83 and reduces V/A-ILR to 0.268/0.327, while introducing only minor changes in Face-Sim and T-Sim. Without progressive training, coverage yields a stronger fidelity–selectivity trade-off, whereas the complete training strategy substantially improves identity preservation, synchronization, speech accuracy, and over-

Table 2: Quantitative ablation study on OC-Bench. We progressively integrate proposed modules to evaluate their individual contributions and further isolate individual module effects in the full setting. **Bold** and underline denote the best and second-best results, respectively.

OCF	SA-MRoPE	SARR		Prog. Train.	Identity Preservation		AV-Sync	Output Quality			Binding Reliability	
		Score	Coverage		Face-Sim/Cons↑	T-Sim↑	Sync-C/D↑↓	IQ↑	WER↓	PQ↑	BA↑	V/A-ILR↓
					0.691 / 0.932	0.402	5.418 / 8.945	0.591	0.346	3.62	0.56	0.556 / 0.468
✓					0.725 / 0.948	0.442	5.512 / 8.645	0.612	0.295	3.82	0.63	0.584 / 0.482
✓	✓				0.745 / 0.956	0.465	5.765 / 8.235	0.628	0.252	3.98	0.68	0.512 / 0.415
✓	✓	✓			0.748 / 0.959	0.468	5.772 / 8.215	0.625	0.245	4.02	0.71	0.465 / 0.362
✓	✓	✓	✓		0.741 / 0.954	0.461	5.768 / 8.224	0.621	0.251	3.99	0.70	0.428 / **0.325**
✓		✓	✓	✓	0.765 / 0.962	0.472	5.925 / 8.114	0.641	0.224	4.15	0.75	0.415 / 0.378
✓	✓			✓	**0.819 / 0.979**	**0.518**	6.155 / 7.912	0.652	0.175	4.31	0.76	0.458 / 0.425
✓	✓	✓		✓	0.816 / 0.976	0.516	6.182 / 7.895	0.654	0.162	4.32	0.81	0.324 / 0.354
✓	✓	✓	✓	✓	0.813 / 0.977	0.515	**6.238 / 7.818**	**0.655**	**0.155**	**4.33**	**0.83**	**0.265** / 0.327



Figure 5: Qualitative ablation study of proposed modules and strategies.

all output quality. These results indicate complementary roles for semantic association, selective reference routing, and progressive training, which are further supported by the qualitative comparisons in Fig. 5.

4.4 User Study and Analysis.

We conduct a blinded user study with 20 ratings per eligible OC-Bench case using criterion-specific candidate sets. All methods in Table 1 are compared for visual quality and image-reference fidelity; seven audio-video generation pipelines are compared for ***Audio–Video Quality***, covering perceptual quality and cross-modal alignment. Only MiniMax-H3, DreamID-Omni, and Omni-Customizer receive both image and voice references in the evaluated configuration, so ***Joint Identity Preservation*** and ***Identity Binding Reliability*** are compared among them on applicable cases. For each criterion, we report the percentage of eligible case–rater votes won within its candidate set. Omni-Customizer achieves win rates of 32.5%, 44.5%, and 46.2% on Audio–Video Quality, Joint Identity Preservation, and Identity Binding Reliability, respectively. More details are provided in Appendix D.

5 Conclusion

Omni-Customizer combines OCF and SA-MRoPE to jointly encode text, visual, and audio identity conditions and establish subject-level cross-modal associations, together with SARR to selectively regulate where and how broadly each reference can directly influence the generated content, reducing face duplication and timbre leakage under incomplete reference coverage. A progressive three-stage training strategy further improves training stability and robustness in complex multi-subject settings with heterogeneous reference availability. Experiments on OC-Bench demonstrate strong video and audio quality, robust appearance-and-voice preservation, and faithful identity binding across complex multi-subject settings. Scaling to higher resolutions and longer sequences with long-term identity consistency remains future work.

AI USE STATEMENT

We used AI tools for data curation and reference construction. Qwen3.5-397B-A17B assisted reference-image quality review, and Qwen3-VL-Embedding-8B assisted reference retrieval. Qwen3-ASR and Qwen3-ForcedAligner produced speech transcripts and timings, with FunASR used as a cross-check. Gemini 2.5 Pro generated structured captions and proposed associations among subjects, reference images, and utterances. LongCat-AudioDiT, Dots-TTS, and CosyVoice3 generated candidate speaker-conditioned reference audios with lexical content different from the target speech; selected clips were used as conditioning data, not as evaluated model outputs. We used generative AI assistants to polish manuscript text, improve figure and table presentation.

ETHICS STATEMENT

This work uses human-centric videos from OpenHumanVid, OpenS2V-5M, TalkVid, and an in-house collection, together with a user study in which human raters assess generated outputs (Appendix D). All source data are accessed and used only for research purposes under the applicable licenses, terms, or authorizations permitting such use. Because the source videos and references may contain identifiable faces and voices, we restrict their use to the research objectives of this work and do not assume that research access grants redistribution rights. High-fidelity appearance-and-voice customization may raise privacy and consent concerns and can be misused for impersonation or deceptive media. Accordingly, any release of source-derived media, identity references, generated examples, or model assets will be separately reviewed for source-specific permissions, privacy risks, and potential misuse, and restricted where redistribution is not permitted.

REPRODUCIBILITY STATEMENT

The model formulation, Omni-Context Fusion, SA-MRoPE, Subject-Aware Reference Routing, and training objectives are described in Sec. 3; the three-stage schedule is in Sec. 3.4, and optimization settings are in Sec. 4.1. Secs. 2 and 4.1 describe the data pipeline, OC-Bench, and evaluation metrics. Appendix B details source filtering, multimodal annotation, and reference construction; Appendix C gives backbone and module implementation details and binding-metric definitions; and Appendix D specifies the human-evaluation protocol.

REFERENCES

Reza Abbasi, Ali Nazari, Aminreza Sefid, Mohammadali Banayeeanzade, Mohammad Hossein Rohban, and Mahdieh Soleymani Baghshah. Clip under the microscope: A fine-grained analysis of multi-object representation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9308–9317. IEEE, 2025.

Keyu An, Yanni Chen, Zhigao Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Bo Gong, Xiangang Li, Yabin Li, et al. Fun-asr technical report. *arXiv preprint arXiv:2509.12508*, 2025.

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, et al. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025a.

Liyang Chen, Tianxiang Ma, Jiawei Liu, Bingchuan Li, Zhuowei Chen, Lijie Liu, Xu He, Gen Li, Qian He, and Zhiyong Wu. Humo: Human-centric video generation via collaborative multi-modal conditioning. *arXiv preprint arXiv:2509.08519*, 2025b.

Qin Chen, Yingjie Chen, Shilun Lin, Cai Xing, Binxin Yang, Long Zhou, Qixin Yan, Wenjing Wang, Dingming Liu, Hao Liu, Chen Li, and Jing Lyu. Identity as presence: Towards appearance and voice personalized joint audio-video generation, 2026a. URL https://arxiv.org/abs/2603.17889.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

Shunian Chen, Hejin Huang, Yexin Liu, Zihan Ye, Pengcheng Chen, Chenghao Zhu, Michael Guan, Rongsheng Wang, Junying Chen, Jianye Hou, et al. Talkvid: A large-scale diversified dataset for audio-driven talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3492–3500, 2026b.

Wei-Ting Chen, Gurunandan Krishnan, Qiang Gao, Sy-Yen Kuo, Sizhou Ma, and Jian Wang. Dsl-fiqa: Assessing facial image quality via dual-set degradation learning and landmark-guided transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2931–2941, 2024.

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28901–28911, 2025.

Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pp. 251–263. Springer, 2016.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Aviad Dahan, Moran Yanuka, Noa Kraicer, Lior Wolf, and Raja Giryes. ID-LoRA: Identity-driven audio-video personalization with in-context LoRA, 2026. URL https://arxiv.org/abs/2603.10256.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, et al. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*, 2025.

Zhengcong Fei, Debang Li, Di Qiu, Jiahua Wang, Yikun Dou, Rui Wang, Jingtao Xu, Mingyuan Fan, Guibin Chen, Yang Li, et al. Skyreels-a2: Compose anything in video diffusion transformers. *arXiv preprint arXiv:2504.02436*, 2025.

Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Dechao Meng, Jinwei Qi, Penchong Qiao, Zhen Shen, Yafei Song, et al. Wan-s2v: Audio-driven cinematic video generation. *arXiv preprint arXiv:2508.18621*, 2025.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15180–15190, 2023.

Jia Guo, Jiankang Deng, et al. InsightFace: 2d and 3d face analysis project. https://github.com/deepinsight/insightface, 2018. Open-source face analysis toolbox.

Xu Guo, Fulong Ye, Qichao Sun, Liyang Chen, Bingchuan Li, Pengze Zhang, Jiawei Liu, Songtao Zhao, Qian He, and Xiangwang Hou. Dreamid-omni: Unified framework for controllable human-centric audio-video generation. *arXiv preprint arXiv:2602.12160*, 2026.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Teng Hu, Zhentao Yu, Guozhen Zhang, Zihan Su, Zhengguang Zhou, Youliang Zhang, Yuan Zhou, Qinglin Lu, and Ran Yi. Harmony: Harmonizing audio and video generation through cross-task synergy. *arXiv preprint arXiv:2511.21579*, 2025a.

Teng Hu, Zhentao Yu, Zhengguang Zhou, Sen Liang, Yuan Zhou, Qin Lin, and Qinglin Lu. Hunyuancustom: A multimodal-driven architecture for customized video generation. *arXiv preprint arXiv:2505.04512*, 2025b.

Daili Hua, Xizhi Wang, Bohan Zeng, Xinyi Huang, Hao Liang, Junbo Niu, Xinlong Chen, Quanqing Xu, and Wentao Zhang. Vabench: A comprehensive benchmark for audio-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23345–23355, 2026.

Yuzhou Huang, Ziyang Yuan, Quande Liu, Qiulin Wang, Xintao Wang, Ruimao Zhang, Pengfei Wan, Di Zhang, and Kun Gai. Conceptmaster: Multi-concept video customization on diffusion transformer models without test-time tuning. *arXiv preprint arXiv:2501.04698*, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17191–17202, 2025.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5148–5157, 2021.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Hui Li, Mingwang Xu, Yun Zhan, Shan Mu, Jiaye Li, Kaihui Cheng, Yuxuan Chen, Tan Chen, Mao Ye, Jingdong Wang, et al. Openhumanvid: A large-scale high-quality dataset for enhancing human-centric video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7752–7762, 2025a.

Maomao Li, Zhen Li, Kaipeng Zhang, Guosheng Yin, Zhifeng Li, and Dong Xu. OmniCustom: Sync audio-video customization via joint audio-video generation model, 2026a. URL `https://arxiv.org/abs/2602.12304`.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, et al. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026b.

Zhaoyang Li, Dongjun Qian, Kai Su, Qishuai Diao, Xiangyang Xia, Chang Liu, Wenfei Yang, Tianzhu Zhang, and Zehuan Yuan. Bindweave: Subject-consistent video generation via cross-modal integration. *arXiv preprint arXiv:2510.00438*, 2025b.

Shi Lian, Changtao Li, Bohan Li, Hankun Wang, Da Zheng, Junfeng Tian, Yufeng Ma, Colin Zhang, and Kai Yu. dots. tts technical report. *arXiv preprint arXiv:2606.07080*, 2026.

Feng Liang, Haoyu Ma, Zecheng He, Tingbo Hou, Ji Hou, Kunpeng Li, Xiaoliang Dai, Felix Juefei-Xu, Samaneh Azadi, Animesh Sinha, et al. Movie weaver: Tuning-free multi-concept video personalization with anchored prompts. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13146–13156, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Kai Liu, Wei Li, Lai Chen, Shengqiong Wu, Yanhao Zheng, Jiayi Ji, Fan Zhou, Jiebo Luo, Ziwei Liu, Hao Fei, et al. Javisdit: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. *arXiv preprint arXiv:2503.23377*, 2025a.

Kaiqi Liu, Yunyao Mao, Ziqi Cai, Zheng Geng, Jing Wang, Qiulin Wang, Xintao Wang, Pengfei Wan, Kun Gai, Shuchen Weng, and Boxin Shi. MAVIN: Multi-shot audio-visual generation with customized narrative control, 2026. URL https://arxiv.org/abs/2606.29473.

Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Gen Li, Siyu Zhou, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14951–14961, 2025b.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22139–22149. IEEE, 2024.

Chetwin Low, Weimin Wang, and Calder Katyal. Ovi: Twin backbone cross-modal fusion for audio-video generation. *arXiv preprint arXiv:2510.01284*, 2025.

Shuailei Ma, Jiaqi Liao, Xinyang Wang, Jingjing Wang, Chaoran Feng, Zijing Hu, Chong Bao, Zichen Xi, Yuqi Gan, Weisen Wang, et al. Scaling mixture-of-experts video pretraining for embodied intelligence. *arXiv preprint arXiv:2607.07675*, 2026.

MiniMax. MiniMax H3: An open model breaking the boundaries between tasks and modalities. https://www.minimax.io/blog/minimax-h3, July 2026. Official technical blog.

OpenMOSS Team, Donghua Yu, Mingshu Chen, Qi Chen, Qi Luo, Qianyi Wu, Qinyuan Cheng, Ruixiao Li, Tianyi Liang, Wenbo Zhang, et al. Mova: Towards scalable and synchronized video-audio generation. *arXiv preprint arXiv:2602.08794*, 2026.

Panwang Pan, Jingjing Zhao, Yuchen Lin, Chenguo Lin, Chenxin Li, Hengyu Liu, Tingting Shen, and Yadong Mu. Id-crafter: Vlm-grounded online rl for compositional multi-subject video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36627–36637, 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

Haoyuan Shi, Yunxin Li, Nanhao Deng, Zhenran Xu, Xinyu Chen, Longyue Wang, Baotian Hu, and Min Zhang. Msvbench: Towards human-level evaluation of multi-shot video generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 24034–24058, 2026a.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, et al. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*, 2026b.

Yuchen Sun, Qian Yang, Jun Wang, Detai Xin, Guoqiao Yu, Guanglu Wan, and Qi Jia. Prism-bench: An audio-centric diagnostic benchmark for text-to-audio-video generation. *arXiv preprint arXiv:2609.04867*, 2026.

Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 3927–3935, 2021.

Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, et al. Audiobox: Unified audio generation with natural language prompts. *arXiv preprint arXiv:2312.15821*, 2023.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Duomin Wang, Wei Zuo, Aojie Li, Ling-Hao Chen, Xinyao Liao, Deyu Zhou, Zixin Yin, Xili Dai, Daxin Jiang, and Gang Yu. Universe-1: Unified audio-video generation via stitching of experts. *arXiv preprint arXiv:2509.06155*, 2025.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Detai Xin, Shujie Hu, Chengzuo Yang, Chen Huang, Guoqiao Yu, Guanglu Wan, and Xunliang Cai. Longcat-audiodit: High-fidelity diffusion text-to-speech in the waveform latent space. *arXiv preprint arXiv:2603.29339*, 2026.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

Shenghai Yuan, Xianyi He, Yufan Deng, Yang Ye, Jinfa Huang, Bin Lin, Jiebo Luo, and Li Yuan. Opens2v-nexus: A detailed benchmark and million-scale dataset for subject-to-video generation. *arXiv preprint arXiv:2505.20292*, 2025.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, et al. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114, 2022.

Ziwei Zhou, Zeyuan Lai, Rui Wang, Yifan Yang, Yuqing Yang, Qi Dai, Lili Qiu, and Chong Luo. Avgen-bench: A task-driven benchmark for multi-granular evaluation of text-to-audio-video generation. In *Forty-third International Conference on Machine Learning*, 2026.

A Related Work

Video foundation models. Recent advances in diffusion transformers and large-scale video training have substantially improved the fidelity, motion quality, and prompt following of open-domain video generation. Representative open-source video foundation models include CogVideoX (Yang et al., 2025), HunyuanVideo (Kong et al., 2024), Wan (Wan et al., 2025), Open-Sora (Zheng et al., 2024), SkyReels-V2 (Chen et al., 2025a), and more recent large-scale video generators such as LingBot-Video (Ma et al., 2026). Building on these strong generative priors, subject-driven video generation further introduces reference images or videos to preserve user-specified appearances and identities. However, these video-centric models primarily focus on the visual modality and do not natively address the joint preservation and generation of appearance, voice identity, speech content, and their

cross-modal correspondence, which is the focus of our work.

Joint audio-video generation. Recent works have moved beyond cascaded video-to-audio pipelines toward native joint audio-video generation. JavisDiT (Liu et al., 2025a) introduces a joint diffusion transformer with hierarchical spatiotemporal priors for synchronized audio-video synthesis. Ovi (Low et al., 2025) adopts twin video and audio DiT backbones with blockwise bidirectional cross-modal interaction, while LTX-2 (HaCohen et al., 2026) employs asymmetric video and audio streams coupled through bidirectional cross-attention. UniVerse-1 (Wang et al., 2025) constructs a unified generator by stitching pretrained video and audio experts, and MOVA (OpenMOSS Team et al., 2026) scales joint generation through a sparse mixture-of-experts architecture. Harmony (Hu et al., 2025a) further studies synchronization during joint diffusion through cross-task training and explicit synchronization-oriented interaction and guidance. Most recently, MiniMax-H3 (MiniMax, 2026) provides a unified audio-video foundation model with native reference-conditioned Ref2VA generation. These models establish increasingly strong joint generation backbones, but their primary objective is general-purpose audio-video synthesis rather than fine-grained multi-subject appearance-and-voice customization.

Video and audio customization. Identity-preserving video generation has progressed from single-subject personalization to open-domain and multi-subject customization. Phantom (Liu et al., 2025b), Movie Weaver (Liang et al., 2025), ConceptMaster (Huang et al., 2025), and BindWeave (Li et al., 2025b) study reference-conditioned subject preservation through cross-modal representations, anchored prompts, concept decoupling, or pre-generation reference–text interaction. ID-Crafter (Pan et al., 2026) further addresses compositional multi-subject generation through hierarchical identity-preserving attention and VLM-guided optimization. These methods focus primarily on visual identity customization without jointly generating the corresponding personalized voices. More recent studies extend identity customization to joint audio-video generation. DreamID-Omni (Guo et al., 2026) combines structured subject descriptions with generator-side Syn-RoPE for reference-conditioned audio-video generation; OmniCustom (Li et al., 2026a) augments Ovi with image and audio reference branches together with reference-conditioned flow contrast; Identity-as-Presence (Chen et al., 2026a) introduces shared identity representations, subject-anchored descriptions, and decoupled reference attention; and ID-LoRA (Dahan et al., 2026) adapts an LTX-based joint generator for appearance-and-voice personalization. MAVIN (Liu et al., 2026) further extends character-conditioned joint generation to multi-shot narratives with optional character references and identity-aware reference propagation. Our work follows this joint-customization setting but separates two roles: OCF associates heterogeneous appearance and voice references with their semantic subjects before denoising, whereas SARR acts on the intra-modal target-query-to-reference-key logits during denoising. In particular, SARR combines target-position subject scores with a predicted coverage budget to limit the direct access range of each available reference under incomplete subject or modality coverage.

Evaluation and benchmarks. The rapid development of generative video models has also motivated increasingly fine-grained evaluation protocols. General-purpose benchmarks such as VBench (Huang et al., 2024) and EvalCrafter (Liu et al., 2024) evaluate visual quality, temporal consistency, motion, and text alignment, but are not designed specifically for reference-conditioned identity preservation. OpenS2V-Nexus (Yuan et al., 2025) introduces OpenS2V-Eval for subject-to-video generation, emphasizing subject consistency, natural appearance, and text relevance over diverse reference subjects. For multi-shot generation, MSVBench (Shi et al., 2026a) evaluates hierarchical story and shot-level conditions together with cross-shot consistency and narrative quality. Evaluation of native audio-video generation has recently been studied by VABench (Hua et al., 2026) and AVGen-Bench (Zhou et al., 2026), which extend evaluation beyond visual fidelity to audio quality, audio-video synchronization, cross-modal semantics, and fine-grained controllability. PRISM-Bench (Sun et al., 2026) further provides an audio-centric diagnostic protocol covering speech, music, and sound under both on-screen and off-screen conditions. For joint identity customization, DreamID-Omni introduces IDBench-Omni (Guo et al., 2026), whose R2AV setting evaluates reference-conditioned appearance and voice generation. These benchmarks provide complementary views of general video quality, subject consistency, multi-shot coherence, and joint audio-video quality; our OC-Bench focuses specifically on multi-subject appearance–voice assignment, including dialogue, silent subjects, off-screen speakers, and incomplete identity-reference conditions.

B HUMAN-CENTRIC DATA CONSTRUCTION

This section provides the implementation details of the data construction pipeline summarized in Sec. 2. We distinguish the filtered target clip pool from the broader reference-candidate pool: target clips satisfy the media and shot constraints below, whereas identity-matched reference images may be retrieved from the full source collections. The pipeline produces a target audio–video clip, a structured multimodal prompt, and identity-aligned image and audio references for each valid subject. We use conservative filtering throughout: records with uncertain identity correspondence, unreliable speech attribution, or invalid reference media are excluded rather than assigned a potentially incorrect label.

B.1 DATA SOURCING AND PRE-FILTERING

Data sources. We collect human-centric videos from OpenHumanVid (Li et al., 2025a), OpenS2V-5M (Yuan et al., 2025), TalkVid (Chen et al., 2026b), and an in-house video collection. We first remove unavailable or corrupted media and then apply the quality scores supplied by each source when such metadata is available.

Clip filtering. We retain clips with a resolution of at least 480p, a frame rate of at least 24 FPS, and a duration between 3 and 10 seconds. Shot-boundary detection is used to exclude clips containing a scene transition, so every retained sample is a single-shot clip. We also discard clips whose audio track is missing or cannot be decoded. These criteria are applied consistently across sources before identity annotation or reference construction.

Data scale. The filtered target pool contains 1,027,513 single-shot clips, comprising 385,317 single-subject clips and 642,196 multi-subject clips. These are counts of target clips rather than derived reference pairings or training exposures.

B.2 FOUR-STAGE CURATION PIPELINE

B.2.1 STAGE 1: REFERENCE IMAGE ACQUISITION

The goal of this stage is to construct a diverse set of high-quality reference images for each identity while avoiding references that simply reproduce the target frame or its surrounding scene. Candidate frames are drawn from the broader source collections rather than only from the filtered target clips.

(i) Face tracking. We sample frames from the source videos at 1 FPS and detect faces with InsightFace (Guo et al., 2018). Detections are associated over time using spatial continuity and ArcFace (Deng et al., 2019) appearance similarity. This produces face tracklets and prevents repeated detections of the same person within a clip from being treated as different identities.

(ii) Face clustering. For each tracklet, we aggregate its ArcFace embeddings into a track-level identity representation. Tracklets are then conservatively clustered across clips according to identity similarity. Candidate merges with competing or inconsistent identity evidence are discarded, since an incorrect merge would propagate directly into both image-reference pairing and speaker attribution.

(iii) Quality filtering. We use DSL-FIQA (Chen et al., 2024) to remove low-quality face crops and apply an additional visual-quality review with Qwen3.5-397B-A17B (Qwen Team, 2026). Images with severe blur, occlusion, extreme truncation, or insufficient facial detail are rejected. The remaining images are verified again with ArcFace before they enter the reference-candidate pool.

(iv) Target–reference pairing. A valid reference must have high ArcFace similarity to the target identity. Whenever possible, it is selected from a different clip or a non-overlapping temporal region of the source video. Among identity-consistent candidates, we prefer images with lower full-frame similarity to the target clip, measured with Qwen3-VL-Embedding-8B (Li et al., 2026b). This pairing rule preserves identity while reducing shared background, pose, and scene content that could otherwise encourage a copy shortcut. If no candidate provides reliable identity evidence, the image reference is marked unavailable.

(v) Reference diversity. When several valid candidates remain, we retain references that cover different viewpoints, expressions, face scales, and backgrounds instead of selecting near-duplicate

frames. Each selected image is resized with its aspect ratio preserved; the short side is padded symmetrically with white pixels to match the target resolution expected by the model.

B.2.2 Stage 2: ASR Annotation and Alignment

The ASR annotations serve two purposes: they provide speech content and timing to Gemini for structured multimodal annotation, and they identify candidate speech intervals for subsequent reference-audio construction.

We first transcribe each clip with Qwen3-ASR and refine the utterance boundaries using Qwen3-ForcedAligner (Shi et al., 2026b). Each resulting record contains the transcript, language, and start and end times. We independently run FunASR (An et al., 2025) as a cross-check. After normalizing punctuation and formatting, records with substantial disagreement in content, language, or temporal coverage are flagged for reprocessing or rejected. Forced alignment is used only to refine speech boundaries; it is not treated as evidence of speaker identity. Likewise, an off-screen utterance is retained when its content and timing are reliable, while its identity attribution is deferred to the multimodal annotation stage.

B.2.3 Stage 3: Gemini-Based Structured Annotation

Gemini 2.5 Pro (Comanici et al., 2025) receives the target audio–video clip, the candidate reference images, and the aligned transcript. It produces a structured annotation with two complementary parts. The subject definition assigns a local anchor such as `<subject 1>` to each person and records which candidate reference image and utterance belong to that anchor. The detailed description associates these subject anchors with their appearance, actions, visible speaking intervals, dialogue, and relevant acoustic context. Spoken content follows the dialogue-span format introduced in Sec. 3.1; the detected language is stored as record metadata rather than inserted as an additional prompt tag.

The annotation does not assume that every utterance belongs to a visible person. Off-screen speech is preserved and assigned to a subject only when the audio–visual and contextual evidence is sufficient; otherwise the attribution is marked uncertain and is not used as identity supervision. A deterministic validation pass checks that every subject anchor resolves to its declared assets, dialogue timestamps fall within the clip, and the same local anchor is used consistently throughout the record.

B.2.4 Stage 4: Reference Audio Acquisition

Ground-truth speech extraction. We use the aligned transcript together with Gemini's subject–utterance association to locate a real speech segment for each valid speaker. A donor segment must contain a single attributable speaker, a complete utterance, and no overlapping speech. Its boundaries are clipped to the valid waveform range, a short context margin is retained to avoid cutting phonemes, and excessive leading or trailing silence is removed. Segments with truncation, clipping, unreliable attribution, or insufficient clean speech are rejected. Off-screen speech may be used when its speaker identity is supported by the structured annotation rather than inferred only from visual presence.

Multi-model synthesis and selection. The retained real segment and its transcript form the voice-cloning condition. To reduce lexical copying from the target clip, the synthesis text uses different content in the same language. We generate candidates with several open-source voice-cloning models, including LongCat-AudioDiT, Dots-TTS, and CosyVoice3 (Xin et al., 2026; Lian et al., 2026; Du et al., 2025), and apply the same speaker-similarity, content-consistency, and audio-quality checks to every candidate. When several candidates pass all checks, we select the one with the highest WavLM-Large (Chen et al., 2022) similarity to the corresponding real donor speech.

Two-stage output filtering. We apply two complementary filters after synthesis. First, a *content-consistency filter* re-transcribes the generated waveform and compares it with the requested synthesis text. It rejects hallucinated, missing, repeated, or unintended words and language mismatches. Second, an *acoustic-purity filter* rejects outputs containing prominent background sounds, music, singing, opera-style vocalization, clipping, or long low-energy regions. Because these heterogeneous failures are not reliably summarized by a single score, ambiguous outputs are conservatively rejected after an additional listening check. We retain a 5-second window only when it contains a

complete synthesized utterance with minimal silence; candidates for which this is not possible are discarded.

Final record validation. The four stages are joined through the sample identifier and local subject anchor rather than through positional order alone. Before a record is used for training, we verify media readability, timestamp bounds, subject–image and subject–audio correspondence, prompt-to-asset consistency, and explicit reference-availability flags. Any record that fails these checks is corrected and revalidated or excluded from identity-specific supervision.

C Implementation Details

This section specifies the backbone, OCF, and SARR implementations used by Omni-Customizer, followed by the implementation of our binding-reliability metrics. All model dimensions refer to the score model unless stated otherwise.

C.1 Ovi Backbone

Dual-stream denoiser. We initialize the score model from Ovi (Low et al., 2025), which contains a 30-block video tower and a 30-block audio tower. Both towers use a hidden width of $d = 3072$, 24 attention heads with a head dimension of 128, and an FFN width of 14336. Within each block, the modality-specific hidden states first undergo intra-modal self-attention. The text-conditioned branch and the cross-modal branch are then evaluated within the same fusion module, combined, and passed through its output projection before the FFN. They are therefore not two successive residual blocks. The video-to-audio and audio-to-video projections are distinct, preserving the asymmetric token structures of the two streams. SARR is applied only to the target-query to reference-key logits of the intra-modal self-attention; it adds no routing bias to the text-conditioned or audio–video interaction branches.

Table 3: Architecture and parameter counts of the score model. Codec and text-encoder parameters are excluded. Counts are obtained directly from the corresponding module shapes.

Component	Configuration	Parameters
Video tower	30 blocks, $d = 3072$	5.566B
Audio tower	30 blocks, $d = 3072$	6.094B
Bidirectional A–V projections	included above	1.133B
Ovi score model	60 tower blocks	11.661B
OCF	8 blocks	1.009B
Shared SARR projections	$W_{\mathrm{anc}}, \{W_c^m\}_{m\in\{v,a\}}$	9.45M
Ovi + OCF	subtotal before SARR	12.669B

Audio representation. The audio stream follows the Ovi audio pipeline based on the MMAudio codec (Cheng et al., 2025). Input waveforms are resampled to 16 kHz and converted with a 1024-point STFT, a hop size of 256, and an 80-bin log-mel front end. The one-dimensional VAE produces 20-channel latents at approximately 31.25 latent tokens per second; for example, a 5-second waveform yields 157 audio tokens after boundary handling. The VAE decoder reconstructs a mel representation, which is converted to a 16 kHz waveform by BigVGAN. The audio VAE and vocoder contain approximately 171.66M and 112.23M parameters, respectively, and are not included in Table 3.

C.2 Omni-Context Fusion

Let L_{txt} be the number of text tokens, let R_v^k and R_a^k be the numbers of visual and audio reference tokens for subject k, and let $m \in \{v, a\}$ index the two denoising streams. After modality-specific input projections, the shared OCF is called once for each stream and receives

$$\mathbf{H}_0^m = [c_{\mathrm{txt}}^m \oplus c_v^1 \oplus \cdots \oplus c_v^N \oplus c_a^1 \oplus \cdots \oplus c_a^N] \in \mathbb{R}^{B\times L_{\mathrm{ocf}}\times 3072}, \tag{5}$$

where $L_{\mathrm{ocf}} = L_{\mathrm{txt}} + \sum_k (R_v^k + R_a^k)$. OCF contains $L = 8$ pre-normalized Transformer blocks whose parameters are shared by the two calls. The superscript m identifies the text context used

by the corresponding denoising tower: the two calls share the OCF parameters and the organized reference set, but their output text tensors are stream-specific. Thus, "shared conditioning" in the main paper refers to the shared condition encoder and subject organization, not to an assumption that $\widetilde{c}^{v}_{\mathrm{txt}}$ and $\widetilde{c}^{a}_{\mathrm{txt}}$ are identical. For block ℓ, we use

$$\mathbf{U}^m_\ell = \mathbf{H}^m_{\ell-1} + W^\ell_o \,\mathrm{Attn}\big(\mathrm{LN}(\mathbf{H}^m_{\ell-1}); \text{SA-MRoPE}\big)\,, \tag{6}$$

$$\mathbf{H}^m_\ell = \mathbf{U}^m_\ell + W^\ell_2 \,\mathrm{GELU}\big(W^\ell_1 \,\mathrm{LN}(\mathbf{U}^m_\ell)\big)\,. \tag{7}$$

The attention uses 24 heads of width 128, and the FFN expands from 3072 to 14336 before projecting back to 3072. We zero-initialize the attention output projection W^ℓ_o and the second FFN projection W^ℓ_2. Consequently, every OCF block is an identity map at initialization, while the remaining attention and FFN weights use their standard initialization.

OCF does not write its visual or audio outputs into the DiT residual streams. After the eighth block, it returns only the change at the text positions:

$$\Delta c^m_{\mathrm{txt}} = \mathbf{H}^m_L[:, 1{:}L_{\mathrm{txt}}] - \mathbf{H}^m_0[:, 1{:}L_{\mathrm{txt}}], \qquad \widetilde{c}^m_{\mathrm{txt}} = c^m_{\mathrm{txt}} + \Delta c^m_{\mathrm{txt}}. \tag{8}$$

Each denoising tower then queries its corresponding $\widetilde{c}^m_{\mathrm{txt}}$ through the original text cross-attention layers. Because OCF depends only on the conditioning inputs and not on the flow timestep, we evaluate these two shared-weight calls once for each classifier-free-guidance branch before the denoising loop and cache the resulting text conditioning.

SA-MRoPE coordinate construction. We assign coordinates by scanning the subject-definition portion of the prompt in its actual order. At the declaration of subject k, α_k denotes the running first-axis RoPE coordinate of the last token of its anchor. The next two first-axis coordinates are reserved for that subject's visual and audio reference tokens, and subsequent text resumes after these reservations. Consequently, α_k already includes the reservations introduced by all earlier subject declarations. Reference coordinates are reserved once at the subject declaration; later mentions of the same anchor in the detailed prompt use the backbone's ordinary text-coordinate rule and do not append the references again. SA-MRoPE otherwise retains the backbone's native coordinate convention for text tokens and the spatial or temporal coordinates of the reference tokens.

C.3 Subject-Aware Reference Routing

Scope and placement. For modality $m \in \{v, a\}$, the input to each intra-modal self-attention layer contains target tokens followed by subject-partitioned reference tokens. These are the modality-specific reference inputs appended to the denoiser; they are distinct from OCF's internal non-text states, which are not written into the DiT residual streams. SARR adds a routing bias only to the pre-softmax logits from target queries to reference keys. It does not directly edit target-to-target logits, reference-query rows, the text-conditioned branch, or the bidirectional audio–video branch.

This scope concerns logits and connectivity rather than the final normalized attention weights. Accessible reference keys participate in the same normalization as target keys, so changing the reference logits can also change the normalized target-to-target weights and the resulting target output. SARR therefore constrains the direct target-query-to-reference-key path; it does not assert global conditional independence, since OCF-enhanced text conditioning and the backbone's native audio–video interactions remain active.

Stream-specific anchors and routing scores. Let $\mathcal{T}_k$ denote the text-token positions of the anchor for subject k. Because OCF returns a stream-specific enhanced text context, we use

$$\mathbf{a}^m_k = W_{\mathrm{anc}} \left(\frac{1}{|\mathcal{T}_k|} \sum_{i \in \mathcal{T}_k} \widetilde{c}^m_{\mathrm{txt},i} \right), \qquad m \in \{v, a\}, \tag{9}$$

where the projection W_{anc} is shared by both streams and all DiT blocks. The main paper suppresses the superscript m on $\mathbf{a}^m_k$ for readability. At a routed self-attention layer, the relevance of valid target token u to subject k is

$$p^m_{u,k} = \mathrm{sigmoid}\Big((\mathbf{h}^m_u)^\top \mathbf{a}^m_k / \sqrt{d} \Big)\,. \tag{10}$$

The anchor is computed from the cached conditioning context, whereas $\mathbf{p}_k^m$ is recomputed from the evolving target states at every routed self-attention layer and denoising evaluation. The scalar routing probability is shared across attention heads.

Let $\mathcal{V}_m$ be the valid, non-padding target positions and $N_m = |\mathcal{V}_m|$. A modality-specific head predicts the coverage fraction $\widehat{\gamma}_k^m = \mathrm{sigmoid}(W_c^m \mathbf{a}_k^m)$. With the reference-availability indicator r_k^m, the token budget and accessible set are

$$C_k^m = r_k^m \left\lceil N_m \widehat{\gamma}_k^m \right\rceil, \qquad S_k^m = \mathrm{TopC}\big(\{p_{u,k}^m : u \in \mathcal{V}_m\}, C_k^m\big). \tag{11}$$

The coverage heads W_c^v and W_c^a are modality-specific and shared across DiT blocks. The integer budget is determined by the cached subject condition; the accessible set is recomputed from the layer-specific routing scores.

Routing and coverage supervision. For each subject and modality, let $\Omega_k^m \subseteq \mathcal{V}_m$ contain reliably annotated positive positions. Video positives are obtained by projecting the subject's annotated face track to the latent-token grid, and audio positives by projecting its attributed speech intervals to the latent timeline. For subject k, positions belonging to other annotated subjects are reliable negatives. Padding, unassigned background, non-speech audio, and ambiguous or missing annotations are not automatically treated as BCE negatives. Writing $\mathcal{M}_k^m$ for the union of the reliable positive and negative positions, the routing target is

$$y_{u,k}^m = \mathbf{1}[u \in \Omega_k^m], \qquad u \in \mathcal{M}_k^m. \tag{12}$$

The routing loss averages binary cross-entropy over subjects, modalities, routed layers, and positions in $\mathcal{M}_k^m$.

The coverage target uses a different denominator: it is the fraction of all valid target positions occupied by the positive region,

$$\gamma_k^m = \frac{|\Omega_k^m|}{N_m}. \tag{13}$$

We supervise $\widehat{\gamma}_k^m$ with an ℓ_1 loss when the corresponding reference is available. The complete training objective can be written as

$$\mathcal{L} = \mathcal{L}_{\mathrm{FM}} + \lambda_{\mathrm{route}} \mathcal{L}_{\mathrm{route}} + \lambda_{\mathrm{cov}} \mathcal{L}_{\mathrm{cov}}, \tag{14}$$

where the two auxiliary losses are averaged only over their valid supervised terms, and we set $\lambda_{\mathrm{route}} = \lambda_{\mathrm{cov}} = 0.1$ in all experiments. Separating $\mathcal{M}_k^m$ from $\mathcal{V}_m$ prevents sparsely annotated background positions from becoming unreliable BCE negatives while retaining the intended full-grid meaning of coverage.

The integer rounding and top-C indices are treated as discrete selections, so we do not differentiate through the budget or ranking operation. The selected $\log p_{u,k}^m$ values remain in the attention computation and are differentiable, while the BCE and coverage objectives provide direct supervision for the continuous scores and coverage fractions.

Reference availability and attention bias. The availability flag is supplied by the conditioning record rather than inferred from placeholder content. A white image or silent waveform may remain in the conditioning tensor after reference dropout to preserve the input structure, but $r_k^m = 0$ forces $C_k^m = 0$ and $S_k^m = \varnothing$. Thus, the direct target-to-reference entries for an unavailable reference are masked in every routed layer.

Reference availability and target presence are different concepts. When a reference is available but its subject has no positive target tokens in a modality, the coverage target is zero and the coverage loss pushes its predicted fraction toward zero; any reliable negative positions additionally supervise the routing scores toward zero. Under the ceiling rule in the main paper, however, a strictly empty direct-access set is guaranteed only by $r_k^m = 0$. For an available reference, C_k^m is an upper bound on direct access rather than a hard prediction that the subject must appear.

For target query u and reference key j belonging to subject k, we apply the logit rule from the main paper,

$$\widetilde{\ell}_{u,j}^m = \begin{cases} \ell_{u,j}^m + \log p_{u,k}^m, & r_k^m = 1 \text{ and } u \in S_k^m, \\ -\infty, & \text{otherwise.} \end{cases} \tag{15}$$

No SARR bias is added to the target-to-target entries. The modified and native logits are then passed to the backbone's ordinary attention normalization. This definition states the routing operation without assuming a particular fused-attention kernel or claiming that indirect conditioning paths are removed.

C.4 Inference Configuration

Our inference procedure follows the official Ovi classifier-free-guidance (CFG) configuration. At each denoising step, we evaluate the joint denoiser twice: a positive branch that uses the shared positive prompt and all available references, and a negative branch that uses modality-specific negative prompts and replaces all reference images and audios with white-image and silent-audio placeholders. The corresponding reference-availability flags are set to zero in the negative branch. Each evaluation jointly produces video and audio predictions, so this procedure requires two model forwards rather than four. The positive and negative OCF contexts are computed once before denoising and cached across all sampling steps.

Let $\mathbf{y}_m^+$ and $\mathbf{y}_m^-$ denote the positive- and negative-branch denoiser predictions for modality m. We use the official modality-specific guidance scales,

$$\widehat{\mathbf{y}}_v = \mathbf{y}_v^- + 4.0\big(\mathbf{y}_v^+ - \mathbf{y}_v^-\big), \tag{16}$$

$$\widehat{\mathbf{y}}_a = \mathbf{y}_a^- + 3.0\big(\mathbf{y}_a^+ - \mathbf{y}_a^-\big). \tag{17}$$

The video negative prompt is `jitter, bad hands, blur, distortion`, and the audio negative prompt is `robotic, muffled, echo, distorted`. We also retain Ovi's skip-layer setting `slg_layer: 11`: only the negative CFG branch skips the fusion block at zero-based index 11, i.e., the 12th fusion block, whereas the positive branch evaluates every fusion block.

C.5 Binding-Reliability Metric Implementation

Evaluation units and reference sets. Each benchmark prompt specifies a set of utterances, the intended subject of each utterance, and the reference modalities supplied for every subject. Binding Accuracy (BA) uses the scripted utterance as its evaluation unit and is computed over utterances whose intended subject has both image and audio references. Audio identity leakage is evaluated over utterances whose intended subject has no audio reference, whereas visual identity leakage is evaluated over annotated visual intervals containing a subject without an image reference. For every metric, $\mathcal{R}_m(c)$ contains only the references of modality $m \in \{v, a\}$ that were actually supplied to the generator in case c. A withheld reference is never used as an identity-matching candidate.

Utterance localization. We decode every generated audio track using a common waveform format and obtain its transcript and word-level timestamps with Qwen3-ASR and the corresponding forced aligner (Shi et al., 2026b). We normalize the scripted and recognized text and associate their word sequences using global Levenshtein alignment. For utterance u, the localization recall is

$$\rho_u = \frac{\#\{\text{script words exactly matched to recognized words}\}}{\#\{\text{script words}\}}. \tag{18}$$

We consider u localized when $\rho_u \geq \tau_{\text{loc}}$, with $\tau_{\text{loc}} = 0.5$. Its temporal span is determined by the earliest and latest timestamps of its matched words, with a context margin $\delta = 0.04$ seconds at each boundary. The normalization rules and these localization parameters are shared by all evaluated methods. This procedure does not assume that generated speech reproduces the script verbatim.

Voice and face identity assignment. For a localized utterance, we extract its waveform segment and encode it with the speaker-verification variant of WavLM (Chen et al., 2022). Reference-audio and generated-speech embeddings are unit-normalized. The predicted voice identity is

$$\widehat{k}_u^a = \begin{cases} \arg\max\limits_{j \in \mathcal{R}_a(c)} \cos(e_a(u), e_a(A_j)), & \max\limits_{j \in \mathcal{R}_a(c)} \cos(e_a(u), e_a(A_j)) > \tau_a, \\ \varnothing, & \text{otherwise.} \end{cases} \tag{19}$$

For the visual branch, we detect and track faces throughout the generated video. ArcFace (Deng et al., 2019) embeddings are normalized per sampled frame, averaged within each track, and normalized again to obtain a track-level representation. TalkNet (Tao et al., 2021) provides an active-speaker score for each face track. For utterance u, we average these scores over its localized interval

and select the valid track with the highest mean score. Its identity $\widehat{k}_u^v$ is assigned by thresholded cosine matching against $\mathcal{R}_v(c)$ using the same nearest-reference rule.

We fix the voice and face verification thresholds to $\tau_a = 0.6$ and $\tau_v = 0.2$, respectively. The same thresholds are applied to all evaluated generators, and generated outputs are not used to tune them.

Binding Accuracy. Let k_u be the intended subject of utterance u, and let o_u indicate that it is specified as off-screen. We define

$$C_u^a = \mathbf{1}[\mathrm{loc}(u) \wedge \widehat{k}_u^a = k_u], \tag{20}$$

$$C_u^v = \mathbf{1}[\mathrm{loc}(u) \wedge (o_u \vee \widehat{k}_u^v = k_u)], \tag{21}$$

$$\mathrm{BA} = \frac{1}{|\mathcal{U}_{\mathrm{BA}}|} \sum_{u \in \mathcal{U}_{\mathrm{BA}}} C_u^a C_u^v, \tag{22}$$

where $\mathcal{U}_{\mathrm{BA}}$ is the set of eligible scripted utterances. A visible utterance is correct only when both its generated voice and its actively speaking face match the intended subject; an off-screen utterance is checked through its voice. Localization failure, a missing embedding, absence of a valid active face for a visible utterance, rejection by either identity threshold, and assignment to a wrong subject all count as errors and remain in the denominator.

Audio identity leakage. Let $\mathcal{U}_a^{\varnothing}$ contain scripted utterances whose intended subject has no audio reference. A leakage event occurs when the generated voice is verified as any audio identity that was supplied in the same case:

$$L_u^a = \mathbf{1}\left[\mathrm{loc}(u) \wedge \max_{j \in \mathcal{R}_a(c)} \cos(e_a(u), e_a(A_j)) > \tau_a\right], \tag{23}$$

$$\text{A-ILR} = \frac{1}{|\mathcal{U}_a^{\varnothing}|} \sum_{u \in \mathcal{U}_a^{\varnothing}} L_u^a. \tag{24}$$

An unlocalized utterance remains in the denominator but cannot be declared a leakage event. We therefore additionally compute Audio Character Presence as a diagnostic coverage measure:

$$\text{A-CP} = \frac{1}{|\mathcal{U}_a^{\varnothing}|} \sum_{u \in \mathcal{U}_a^{\varnothing}} \mathbf{1}[\mathrm{loc}(u) \wedge e_a(u) \text{ is valid}]. \tag{25}$$

This companion statistic prevents omitted or unmeasurable speech from being interpreted as successful identity separation.

Visual identity leakage. Visual leakage is evaluated over all valid face tracks, independently of whether a face is speaking. Each track is compared only with the supplied image references and is assigned a referenced identity when its maximum ArcFace similarity exceeds τ_v. Because occlusion or missed detections may split one person into several non-overlapping tracks, we count simultaneous occupancy rather than raw tracks. For reference identity j in annotated visual interval s, let

$$n_{s,j} = \max_{f \in s} \sum_t \mathbf{1}[f \in t \wedge \widehat{k}_t^v = j] \tag{26}$$

be the maximum number of concurrent tracks verified as j. Let $m_{s,j}$ be the number of legitimate visible owners of identity j specified by the benchmark annotation. The excess referenced-identity count is

$$L_s^v = \sum_{j \in \mathcal{R}_v(c)} \max(0, n_{s,j} - m_{s,j}), \tag{27}$$

and Visual-ILR is

$$\text{V-ILR} = \frac{\sum_{s \in \mathcal{S}_v^{\varnothing}} L_s^v}{|\mathcal{V}_v^{\varnothing}|}, \tag{28}$$

where $\mathcal{S}_v^{\varnothing}$ contains the evaluated visual intervals and $\mathcal{V}_v^{\varnothing}$ contains the visible subject occurrences without an image reference. This formulation detects a supplied identity being copied onto an additional character while avoiding duplicate counts caused only by track fragmentation.

We additionally compute Visual Character Presence as a diagnostic coverage measure. Let q_s be the number of subjects required to be visible in interval s and let d_s be the maximum number of concurrent valid face tracks detected in that interval. We compute

$$\text{V-CP} = \frac{1}{|\mathcal{S}_v^{\varnothing}|} \sum_{s \in \mathcal{S}_v^{\varnothing}} \mathbf{1}[d_s \geq q_s]. \tag{29}$$

V-CP is a presence and measurability safeguard, while V-ILR tests whether a supplied visual identity has spread beyond its legitimate owner. Inspecting each ILR together with its corresponding CP distinguishes reduced identity leakage from the degenerate strategy of omitting the unreferenced character. V-ILR specifically measures excess simultaneous occupancy of a supplied identity. A pure replacement that creates no excess occupancy is outside this metric's scope, which is why V-ILR is interpreted together with V-CP rather than as a complete character-recognition score.

D User Study

Evaluation protocol. We obtain 20 independent ratings per eligible OC-Bench case under blinded, randomized presentation. Candidate methods and case eligibility are specified separately for each criterion before viewing generated outputs. The visual panel covers all 13 methods in Table 1. Raters watch the videos without audio and judge visual quality (VQ) on all 200 cases. Visual identity preservation (V-IP) is judged only when the prompt calls for a visible subject with an image reference; raters inspect the prompt and image references and assess the output's appearance fidelity. An omitted or incorrect subject counts as a failure rather than making that model's output ineligible.

The audio–video panel shows generated video with audio. Audio–video quality (AVQ), covering video quality, generated-audio quality, and audiovisual alignment, compares the four Qwen-Image plus JAVG pipelines, MiniMax-H3, DreamID-Omni, and Omni-Customizer on all 200 cases. Joint identity preservation (Joint-IP) and identity binding reliability (IBR) compare only the reference-conditioned systems MiniMax-H3, DreamID-Omni, and Omni-Customizer, each supplied with image and audio references through its corresponding interface. Joint-IP assesses appearance and voice similarity to the available references on cases with an evaluable image–audio reference pair. IBR assesses whether faces and voices are assigned to their intended subjects without cross-subject leakage on eligible multi-subject cases; off-screen speakers are judged from their voice and the prompt rather than requiring a visible speaking face. The four Qwen-Image plus JAVG configurations are not scored for Joint-IP or IBR because they do not receive reference voices in the evaluated pipeline.

Win-rate calculation. For each criterion, a method's win rate is the percentage of eligible case–rater judgments selecting that method from that criterion's candidate set. VQ and AVQ each have 4,000 judgments (200 cases $\times$ 20 raters); V-IP, Joint-IP, and IBR use their respective pre-specified eligible-case counts. Each column sums to 100% over its eligible methods, up to rounding. Because the candidate sets and sometimes the eligible cases differ, percentages must not be compared across columns. The three audio–video columns correspond to the three win rates reported in Sec. 4.4.

Table 4: Blinded user-study win rates (%) on OC-Bench. Candidate methods are determined separately for each criterion; dashes indicate that the evaluated pipeline lacks the required reference-voice interface. V-IP, Joint-IP, and IBR use pre-specified eligible cases. QI denotes the Qwen-Image first-frame cascade.

Visual preference: all methods			*Audio–video preference: AV-generating methods*			
Method	VQ	V-IP	Method	AVQ	Joint-IP	IBR
Phantom	1.2	1.0	QI + Universe-1	4.5	–	–
VACE	1.5	1.3	QI + Ovi	5.4	–	–
Humo	2.1	1.8	QI + MOVA	6.2	–	–
HunyuanCustom	4.8	4.5	QI + LTX2.3	7.7	–	–
Wan2.2-S2V	4.2	4.0	MiniMax-H3	26.2	35.2	34.5
SkyReel-A2	3.5	3.2	DreamID-Omni	17.5	20.3	19.3
QI + Universe-1	1.8	1.5	**Omni-Customizer**	**32.5**	**44.5**	**46.2**
QI + Ovi	2.4	2.2				
QI + MOVA	2.8	2.6				
QI + LTX2.3	4.2	3.8				
MiniMax-H3	25.8	26.5				
DreamID-Omni	14.5	15.2				
Omni-Customizer	**31.2**	**32.4**				